NPS-OR-06-004

NAVAL
POSTGRADUATE
SCHOOL

MONTEREY, CALIFORNIA

Index Policies for Shooting Problems

by

K.D. Glazebrook
C. Kirkbride
H.M. Mitchell
D.P. Gaver
P.A. Jacobs

January 2006

Approved for public release; distribution is unlimited.

Prepared for: Wayne E. Meyer Institute for Systems Engineering
Maritime Domain Protection Group
777 Dyer Road, Bldg. 233
Monterey, CA 93943-5194

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93943-5001

RDML Richard H. Wells, USN
President

Richard Elster
Provost

This report was prepared for and funded by Wayne E. Meyer Institute for Systems Engineering, Maritime Domain Protection Group, 777 Dyer Road, Bldg. 233, Monterey, CA 93943-5194.

This report was prepared by:

K.D. GLAZEBROOK
Professor of Statistics and Operations Research; Lancaster University

C. KIRKBRIDE
Lancaster University

H.M. MITCHELL
University of Newcastle-upon-Tyne

D.P. GAVER
Distinguished Professor of Operations Research

P.A. JACOBS
Professor of Operations Research

Reviewed by:

LYN R. WHITAKER
Associate Chairman for Research
Department of Operations Research

JAMES N. EAGLE
Chairman
Department of Operations Research

Released by:

LEONARD A. FERRARI, Ph.D.
Associate Provost and Dean of Research

REPORT DOCUMENTATION PAGE *Form Approved OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** January 2006	**3. REPORT TYPE AND DATES COVERED** Technical Report

4. TITLE AND SUBTITLE: Index Policies for Shooting Problems

5. FUNDING NUMBERS
BORCD

6. AUTHOR(S) K.D. Glazebrook, C. Kirkbride, H.M. Mitchell, D.P. Gaver, and P.A. Jacobs

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER NPS-OR-06-004

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Wayne E. Meyer Institute for Systems Engineering
Maritime Domain Protection Research Group
777 Dyer Road, Bldg. 233
Monterey, CA 93943-5194
and the Center for Defense Technology and Education for the Military Services Initiative (CDTEMS)

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(maximum 200 words)*

We consider a scenario in which a single Red wishes to shoot at a collection of Blue targets, one at a time, to maximise some measure of return obtained from Blues killed before Red's own (possible) demise. Such a situation arises in various military contexts such as the conduct of air defence by Red in the face of Blue SEAD (suppression of enemy air defences). A class of decision processes called multi-armed bandits has been previously deployed to develop optimal policies for Red in which she attaches a calibrating (Gittins) index to each Blue target and optimally shoots next at the Blue with largest index value. The current paper seeks to elucidate how a range of developments of index theory are able to accommodate features of such problems which are of practical military import. Such features include levels of risk to Red which are policy dependent, Red having imperfect information about the Blues she faces, an evolving population of Blue targets and the possibility of Red disengagement. The paper concludes with a numerical study which both compares the performance of (optimal) index policies to a range of competitors and also demonstrates the value to Red of (optimal) disengagement.

14. SUBJECT TERMS multi-armed bandits, Gitten Indices, suppression of enemy air defense

15. NUMBER OF PAGES
27

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UL

Index policies for shooting problems

K.D. Glazebrook and C. Kirkbride, Department of Management Science,
Management School, Lancaster University, UK,

H.M. Mitchell, School of Mathematics and Statistics,
Newcastle University, UK,

D.P. Gaver and P.A. Jacobs, Department of Operations Research,
Naval Postgraduate School, Monterey, USA.

Abstract

We consider a scenario in which a single Red wishes to shoot at a collection of Blue targets, one at a time, to maximise some measure of return obtained from Blues killed before Red's own (possible) demise. Such a situation arises in various military contexts such as the conduct of air defence by Red in the face of Blue SEAD (suppression of enemy air defences). A class of decision processes called multi-armed bandits has been previously deployed to develop optimal policies for Red in which she attaches a calibrating (Gittins) index to each Blue target and optimally shoots next at the Blue with largest index value. The current paper seeks to elucidate how a range of developments of index theory are able to accommodate features of such problems which are of practical military import. Such features include levels of risk to Red which are policy dependent, Red having imperfect information about the Blues she faces, an evolving population of Blue targets and the possibility of Red disengagement. The paper concludes with a numerical study which both compares the performance of (optimal) index policies to a range of competitors and also demonstrates the value to Red of (optimal) disengagement.

1 Introduction

A multi-armed bandit problem arises when a single key resource (possibly an enemy defence weapon system, here called *Red*) is available for allocation to a fixed collection of projects or "bandits". The latter may be enemy force elements (here called *Blue*) which are attempting to penetrate space and attack assets guarded by Red. These projects evolve sequentially and stochastically while in receipt of service (i.e. while the resource is allocated to them) and obtain state dependent returns as they do so, but remain fixed (and gain nothing) otherwise. Gittins and Jones (1974) elucidated the optimality of *index policies* for certain classes of multi-armed bandit problems. Such policies attach a calibrating *index* to each project, a function of that project's state, and choose at each decision epoch to allocate the resource to whichever project has the largest associated index. See also Gittins (1989). An extensive literature exists outlining a range of extensions and developments of Gittins' classical work while various schemes for index computation have been proposed. See, for example, Whittle (1980), Weber (1992), Katehakis and Veinott (1987) and Bertsimas and Niño-Mora (1996).

Recently, Glazebrook and Washburn (2004) have discussed the utilisation of the multi-armed bandit framework and the associated index policies to develop optimal shooting policies in a military environment. Here the "key resource" is a single shooter (Red) and the "projects" form a fixed collection of targets (Blue). Specifically, think of Red as a ship or other defence system and the Blues as a collection of attackers. Red's goal is to so target the Blues as to maximise the expected number (or value) of kills achieved. Manor and Kress (1997) had previously utilised the theory of multi-armed bandits to analyse a shooting problem in which Red receives incomplete information regarding the outcome of successive shots. If a shot is unsuccessful (the Blue target is not killed) then Red receives no feedback, while if the target *is* killed, that fact is confirmed to Red with probability less than one. Manor and Kress (1997) demonstrate the optimality of a form of index policy (the greedy shooting policy) for their setup.

Consider a military scenario discussed by Barkdoll *et al.* (2002) which is asymmetric between enemy forces. Blue has established air superiority in some region and Red is a surface-to-air missile system (SAM) seeking to disrupt Blue's air campaign. The U.S. Joint Chiefs of Staff uses the term "reactive" or "opportune" suppression of enemy air defences (SEAD). A U.S. Marine Corps Warfighting Publication (2001) gives a background summary of SEAD operations. In Barkdoll *et al.* (2002) every Red shot exposes her to danger from a stand-off Blue shooter. Red attaches a value to every Blue she faces which could, for example, reflect the damage which would be caused should that Blue penetrate her defences. We take Red's goal to be maximisation of the expected value of Blues killed (rendered ineffective) before her own (possible) demise, thus minimising the cost of Blue leakage to (possible) valuable Red targets. Many important features of such situations present a challenge to analysis. Several have gone largely unconsidered in previous work. It is the prime purpose of the current paper to elucidate how a range of developments of index theory are able to accommodate such features. They include the following:

(1) The level of danger to the Red SAM may vary according to the Blue targets she chooses. For example, shooting at longer range Blues puts Red at greater risk to anti-radiation missile (ARM) attack from Blue since a SAM will need to radiate longer to guide the missile to its target. Red should plainly take account of such risks to herself in deciding which Blue of those currently within range, to target next;

(2) Red may have imperfect information regarding the Blues she is facing, including the efficacy of past shots;

(3) The value which Red attaches to a particular Blue may evolve during battle as, for example, Red gains information about it and/or causes it damage. Operational value tends to be dynamic and not well measured in monetary terms;

(4) The Blue targets which Red faces will change over time. Blues currently within range may withdraw (or may penetrate Red's defences) while new Blues may arrive;

(5) It may be that, such is the nature of Blues which Red currently faces that her best option is to disengage (i.e., defer active engagement) thus reducing her current risks in the interests of securing greater gains from attacking Blues which arrive later.

We present models and analyses which illustrate all of the above. In every case, our goal is to provide Red with an appropriate calibration of her options (including disengagement) at all times.

The paper is structured as follows: In Section 2 we present a class of shooting problems incorporating disengagement in which Red faces a collection of Blue SEAD targets which arrive as a group. These problems take the form of *generalised bandits*, a type of multi-armed bandit problem in which a form of reward dependence is induced between the constituent projects or "bandits" via a multiplicatively separable structure. The models were originally introduced by Nash (1980) and subsequently developed by Fay and Walrand (1991), Glazebrook and Greatrix (1995) and Crosbie and Glazebrook (2000a,b). Applications of Nash's model have recently been described by Dumitriu, Tetali and Winkler (2003) and by Katta and Sethuraman (2004). We describe the nature of optimising *index policies* for Red. The approach is illustrated in Section 3 by analyses of three models of independent interest. Model 1 (in Section 3.1) is a Bayesian model in which Red is able to learn about the (true) identity of the Blues she faces as the engagement proceeds. Blues may withdraw while under fire. Model 2 (in Section 3.2) allows for partial/cumulative damage to each Blue target, while Model 3 (in Section 3.3) extends Model 1 in allowing Red to supplement the information she has about the Blues she faces by "looking" (imperfectly) at the most recently targeted Blue after each shot. In Section 4 we discuss the value to Red of disengagement from targeting the Blues currently present. To achieve this we propose two possible models for the future Blues with which Red will be presented as a process extended over time. Both models yield qualitatively similar insights, namely that the greater the opportunity for a surviving Red to secure future gains from Blue kills, the more selective she should be about the targets to be engaged now. Index theory enables us to quantify Red's selectivity precisely. The paper concludes in Section 5 with a numerical study which, *inter alia*, sheds light on the value to Red of (optimal) disengagement.

2 A General Model for a Single Conflict with Disengagement

A Red shooter has to plan a series of engagements with a *finite fixed collection* of N Blues. At each decision epoch $t \in \mathbb{N}$ for which the conflict is still active, a still alive Red will *either* disengage (i.e., suspend active engagement) and claim a return of R_d *or* will shoot at a targeted Blue. In the former case shooting ceases while in the latter Red is exposed to the possibility of being killed herself. Each shot takes a single time unit. Disengagement return R_d may be understood as the future value available to an optimally shooting/disengaging Red. How R_d could be set is discussed in Section 4. Choice of which Blue to attack will depend not only upon which targets appear most vulnerable and likely to yield high returns for Red but also which expose Red to little risk. An engagement may also incorporate, for example, a look by Red to gain information on the state of the Blue targeted after she has delivered her shot. She is assumed to have an infinite supply of shots (e.g., surface to air missiles). Red will make her decisions on the basis of her observational history of past engagements to date. Her goal is the maximisation of expected return received until her own (possible) death. This is formulated as a discounted reward Markov decision problem as follows:

(i) $\boldsymbol{X}(t) = \{X_1(t), X_2(t), \ldots, X_N(t), X_{N+1}(t)\}$ is the state of the system at time $t \in \mathbb{N}$ and $X_j(t)$ is the state of Blue j, $1 \leq j \leq N$. We require that $X_j(t) \in \Omega_j \cup \{\omega_j\}$, where Ω_j is the (countable) space of possible descriptors of Red's knowledge of Blue's status, while $X_j(t) = \omega_j$ indicates that by time t, Red has been killed (i.e., rendered ineffective) during an engagement in which she shot at Blue j, $1 \leq j \leq N$. $X_{N+1}(t)$ is an indicator which takes the value 0 if Red has disengaged prior to t and is 1 otherwise. We assume that $X_{N+1}(0) = 1$;

(ii) At each $t \in \mathbb{N}$ for which $X_{N+1}(t) = 1$ and $X_j(t) \neq \omega_j$, $1 \leq j \leq N$, (i.e. Red is still alive and has not disengaged), Red must choose one of the actions $a_1, a_2, \ldots, a_N, a_{N+1}$. Choice of a_j means that Red's $(t+1)^{\text{st}}$ engagement will target Blue j, $1 \leq j \leq N$. Choice of a_{N+1} indicates Red's disengagement from further shooting. The above are the only $t \in \mathbb{N}$ for which a decision by Red is required;

(iii) If at decision epoch $t \in \mathbb{N}$ action a_j is taken, $1 \leq j \leq N$, then Red observes a change of Blue's state $X_j(t) \to X_j(t+1)$ determined by some Markov law P_j. Note that state space Ω_j may contain some state $\overline{\omega}_j$ indicating that Blue is dead and that a still alive Red knows this. In such cases, both $\overline{\omega}_j$ and ω_j are absorbing states under P_j. Note that when any action a_k is taken at t, then $X_l(t) = X_l(t+1)$, $l \neq k$;

(iv) The expected return achieved by Red when action a_j, $1 \leq j \leq N$, is taken at time $t \in \mathbb{N}$ is $\beta^t R_j\{X_j(t)\}$ where $R_j : \Omega_j \to \mathbb{R}^+$ is bounded and non-negative and $\beta \in (0,1)$ is effectively a discount rate. The non-negative returns determined by function R_j will reflect the operational value to Red of rendering Blue j ineffective. For example, R_j may estimate the damage caused should a still alive Blue j penetrate Red's defences. Discount rate β will be set to reflect military-operational realities. For example, if Red is exposed to threats *other* than those coming from Blue, then β can be taken to be the probability that she survives this external threat for a single unit of time. See also the comments at the end of this section. The return achieved when action a_{N+1} is taken at time $t \in \mathbb{N}$ is $\beta^t R_d$.

A *policy* for Red is a rule for taking actions which takes account of the history of the process (actions taken, states occupied) to date. The theory of stochastic dynamic programming (see, for example, Puterman (1994)) guarantees the existence of an optimal policy which is stationary, deterministic and Markov. However, we can say more. Consider a modification of (i)-(iv) above which allows actions to be taken at *all* $t \in \mathbb{N}$, but which guarantees that no returns are gained beyond time

$$T = \inf\{t;\ X_j(t) = \omega_j \text{ for some } 1 \leq j \leq N \text{ or } X_{N+1}(t) = 0\}. \tag{1}$$

This can be achieved by a modification to (iv) which requires that, should action a_j be taken at general time $t \in \mathbb{N}$ for some $1 \leq j \leq N$, the resulting expected return is

$$\beta^t R_j\{X_j(t)\} \left[\prod_{k=1}^{N} I\{X_k(t) \neq \omega_k\}\right] X_{N+1}(t), \tag{2}$$

while should disengagement action a_{N+1} be taken at $t \in \mathbb{N}$ the expected return is

$$\beta^t R_d \left[\prod_{k=1}^{N} I\{X_k(t) \neq \omega_k\}\right] X_{N+1}(t). \tag{3}$$

In (2) and (3) above, $I(\cdot)$ is an indicator taking the value 1 when the bracketed event is observed. Plainly an optimal policy for any such modification will map directly to an optimal policy for the process in (i)–(iv) above.

Now write ν for a policy for the above modification with $\nu(t)$ for the choice of action made by ν at t. The total expected return under policy ν may be expressed as

$$E_\nu \left(\sum_{t=0}^{\infty} \beta^t R_{\nu(t)} \left\{ X_{\nu(t)}(t) \right\} \left[\prod_{j=1}^{N} I \left\{ X_j(t) \neq \omega_j \right\} \right] X_{N+1}(t) \right). \tag{4}$$

The goal is to find policy ν^* to maximise the expected return in (4). The multiplicatively separable form of the objective means that the above falls within the class of *generalised bandits* introduced by Nash (1980). Moreover, the facts that in our models rewards are non-negative and that the quantities $I\{X_j(t) \neq \omega_j\}$, $1 \leq j \leq N$ and $X_{N+1}(t)$ can only decrease in value as time proceeds imply that we are dealing with cases of Nash's models which are equivalent to semi-Markov versions of Gittins' (1979,1989) classic multi-armed bandits. See, for example, Fay and Walrand (1991). It then follows that there exists an optimal policy for our problem in simple index form. We express our conclusion in Theorem 1, which also utilises the fact that, while the conflict remains active, the disengagement action has fixed index R_d.

Theorem 1 *There exist functions $G_j : \Omega_j \to \mathbb{R}^+$ such that, while the conflict remains active (i.e., prior to T), Red optimally disengages if*

$$R_d > \max_{1 \leq j \leq N} G_j\{X_j(t)\} \tag{5}$$

and otherwise optimally engages any Blue j^* for which

$$G_{j^*}\{X_j(t)\} = \max_{1 \leq j \leq N} G_j\{X_j(t)\}. \tag{6}$$

The indices in (5), (6) are of Gittins type and are computable by a range of algorithms including the "restart-in-x" approach of Katehakis and Veinott (1987). When the state spaces Ω_j, $1 \leq j \leq N$, are finite the "largest-to-smallest" algorithm of Robinson (1982) (equivalently, the adaptive greedy algorithm of Bertsimas and Niño-Mora (1996)) is available. See also Gittins (1989).

To develop $G_j(x)$ for some $x \in \Omega_j$, suppose that at $t = 0$ Blue j is in state x and is then engaged continuously by Red up to some positive integer-valued stopping time τ defined on Blue's state process $\{X_j(t), t \geq 0\}$. We write $R_j(x, \tau)$ for the expected reward earned by Red during $[0, \tau)$ and also write

$$S_j(x, \tau) = 1 - E \left[\beta^\tau I \left\{ X_j(\tau) \neq \omega_j \right\} | X_j(0) = x \right]. \tag{7}$$

To give the quantity in (7) a simple interpretation, suppose that discount rate β has the interpretation given in (iv) above as the probability that Red survives an external threat for a single unit of time. Under suitable independence assumptions, $S_j(x, \tau)$ is then seen to be the probability that Red fails to survive her engagement with Blue during $[0, \tau)$. The index $G_j(x)$ is given by

$$G_j(x) = \max_\tau \left\{ R_j(x, \tau) / S_j(x, \tau) \right\}, \ x \in \Omega_j, \tag{8}$$

and is seen to balance the rewards which Red can earn from Blue j (as expressed by the numerator on the r.h.s. of (8)) in state x against the risks posed (as expressed by the denominator in (8)).

Comments

(a) Military-operational interpretations of discount rate β other than that identified in (iv) above are certainly possible. Suppose, for example, that the Blue force withdraws from the current conflict (at every epoch at which it is still active) with probability ψ. Successive determinations in this regard are independent. Such Blue "withdrawal" may take the form of leakage through Red's defences with a view to inflicting damage on assets under Red protection. The quantity $1 - \psi$ is suitable as a discount rate. Should an external threat to Red *also* exist, then the discount rate should be taken as the product of Red and Blue's "survival" probabilities for a single time unit.

(b) Note that should Ω_j contain some state $\overline{\omega}_j$ as in (iii) above, then if there exists $\epsilon > 0$ such that

$$P_j(x, \omega_j) \geq \epsilon, \ x \in \Omega_j \setminus \{\overline{\omega}_j\}, \ 1 \leq j \leq N, \tag{9}$$

we can take discount rate β to be equal to one in the above. Under (9) it continues to be possible to construct an equivalent decision process in the form of a discounted reward semi-Markov multi-armed bandit.

(c) The model presented in (i)–(iv) above is rich enough to accommodate a range of assumptions about whether and how Blue might withdraw from the conflict. One scenario has Blue j's death triggering withdrawal of the Blue force with some probability $\psi_j, \ 1 \leq j \leq N$. Scenarios in which individuals may withdraw while under fire are considered in Models 1 and 3 of Section 3.

(d) Suppose now that individual alive Blues leave the conflict (possibly breaching Red's defences) while not under fire from Red with probability η. Distinct Blues leave independently of each other. This modification to the above models is technically far-reaching and the optimisation problem for Red now becomes a *restless bandit problem.* Whittle (1988) introduced restless bandits as an extension of Gittins' (1979,1989) multi-armed bandits in which projects can evolve (i.e. targets leave the conflict) when inactive (i.e. not under fire). Restless bandits are likely intractable (see Papadimitriou and Tsitsiklis (1999)) and Whittle proposed a class of *index heuristics* derived from Lagrangian relaxations of the original optimisation problem. For a problem in which this feature of Blue withdrawal is incorporated into a model with discount rate β then a suitable index for Blue j may be inferred from an argument based on pairwise interchanges. For this index, replace the quantity in (7) by

$$\tilde{S}_j(x, \tau) = 1 - E\left[\{\beta(1-\eta)\}^{\tau} I\left\{X_j(\tau) \neq \omega_j\right\} | X_j(0) = x\right]$$

and then develop index $\tilde{G}_j(x)$ as

$$\tilde{G}_j(x) = \max_{\tau} \left\{ R_j(x, \tau) / \tilde{S}_j(x, \tau) \right\}, \ x \in \Omega_j. \tag{10}$$

A policy based on the indices in (10) (used as in the statement of Theorem 1) should perform strongly. See Glazebrook *et al.* (2004).

In Section 3 we illustrate the above by presenting three models, each of which present salient features of combat scenarios.

3 Index policies for a Range of Single Conflicts with Disengagement

Each of the models presented in this section conform to (i)–(iv) of Section 2. Hence a single Red shooter faces N Blues which may do her harm. At every point in the conflict she may shoot at a Blue or decide to suspend active engagement.

3.1 Model 1 – Red learns about the nature of Blue targets

Suppose that Blues come in B types and that Red has imperfect information about the Blues she is facing. Note that "type" designation here may reflect any Blue characteristics which are relevant to determining outcomes as the conflict proceeds. Red's uncertainty about Blue is expressed through N independent prior distributions Π^1, $\Pi^2, \dots,$ Π^N which summarise her beliefs before shooting starts. Hence Π_b^j is the prior probability that Red assigns to the event "Blue number j is of type b", $1 \leq j \leq N,\ 1 \leq b \leq B$.

We assume here that the type of a Blue does not change through the conflict. At each time $t = 0,\ 1,\ 2, \dots$ at which Red is alive she either targets a single Blue *or* disengages from the conflict. The latter option, when taken at time t, yields a return of $\beta^t R_d$. If a Blue is targeted, then conditional upon the event that the Blue concerned is actually of type b, Red has a probability ρ_b of killing Blue while there is a probability θ_b that she herself is killed during the engagement. Observe that both Blue and Red are subject to attrition. Further, there is a probability ϕ_b that a Blue of type b withdraws from the conflict following an unsuccessful shot by Red. Red always has perfect information about whether each Blue is still present and also whether alive or dead. This optimistic assumption is relaxed in Section 3.3. Hence the model calls for the inclusion of state $\overline{\omega}_j$ within Ω_j as mentioned in Section 2(iii) above. All shooting outcomes are assumed independent. Should Red kill a type b Blue with her t^{th} shot then she gains a return $\beta^t R_b$. Red's goal is to maximise the expected return from Blues killed and from disengagement prior to her own destruction. The expectation concerned is taken both with respect to Red's prior beliefs as well as over realisations of the process. Note that, if all θ_b's are (strictly) positive then we are permitted the choice $\beta = 1$ since an appropriate version of the condition in (9) will be met. If, further, $R_b = 1,\ 1 \leq b \leq B$, then Red's goal is the maximisation of the expected number of Blues killed aggregated with any (future) return from disengagement.

A crucial feature of the model concerns Red's capacity to update her beliefs about the Blues she is facing in the light of past engagements by using Bayes' Theorem. In particular, if Blue j has been targeted in n engagements, has not withdrawn and (along with Red) is still alive (note that these are the only event types of relevance for future decision-making) then the posterior distribution $\Pi^{j,n}$ summarising Red's updated beliefs about Blue j is given by

$$\Pi_b^{j,n} = \frac{\Pi_b^j (1-\rho_b)^n (1-\theta_b)^n (1-\phi_b)^n}{\sum_{d=1}^B \Pi_d^j (1-\rho_d)^n (1-\theta_d)^n (1-\phi_d)^n}, \quad 1 \leq b \leq B. \tag{11}$$

For notational simplicity, we shall refer to the denominator in (11) as $D(j,n)$.

This problem may be represented within the general formulation of Section 2 (i)–(iv) as follows:

(i) State space Ω_j is taken to be $\mathbb{N} \cup \{\overline{\omega}_j\} \cup \{*_j\}$, where $*_j$ is the state entered when a still alive Blue j withdraws from the conflict. If $X_j(t) = n \in \mathbb{N}$, then at time t, Blue j has been targeted in n engagements with Red, all of which have been inconclusive (neither killed), and has not withdrawn.

(iii) Should action a_j be chosen at t when $X_j(t) = n$ then, following the resulting engagement a transition to $X_j(t+1)$ occurs according to Markovian law P_j where

$$\begin{aligned} P_j(n, n+1) &= P(\text{neither Red nor Blue } j \text{ killed and Blue remains in the conflict}) \\ &= D(j, n+1)/D(j,n); \end{aligned}$$

$$\begin{aligned} P_j(n, \overline{\omega}_j) &= P(\text{Blue } j \text{ killed but not Red}) \\ &= \sum_{b=1}^{B} \Pi_b^j \rho_b (1-\rho_b)^n (1-\theta_b)^{n+1} (1-\phi_b)^n / D(j,n); \end{aligned}$$

$$\begin{aligned} P_j(n, *_j) &= P(\text{neither Red nor Blue } j \text{ killed and Blue leaves the conflict}) \\ &= \sum_{b=1}^{B} \Pi_b^j \phi_b (1-\rho_b)^{n+1} (1-\theta_b)^{n+1} (1-\phi_b)^n / D(j,n); \end{aligned}$$

and

$$P_j(n, \omega_j) = P(\text{Red killed}) = \sum_{b=1}^{B} \Pi_b^j \theta_b (1-\rho_b)^n (1-\theta_b)^n (1-\phi_b)^n / D(j,n). \tag{12}$$

The expected return gained from a Blue kill in the engagement in (iii) above is given by

$$R_j(n) = \sum_{b=1}^{B} \Pi_b^j R_b \rho_b (1-\rho_b)^n (1-\theta_b)^n (1-\phi_b)^n / D(j,n), \ n \in \mathbb{N}. \tag{13}$$

With the above specifications the index $G_j(n)$, appropriate for Blue j in state $n \in \mathbb{N}$, may be computed straightforwardly from (8). Red's optimal policy is to target next the still alive and still present Blue with maximal index up to the time of her death or the point at which all still alive and still present Blues have index less than the disengagement return R_d. In the latter event, Red disengages. Note that if at time $t = 0$ all Blues have index less than R_d, Red does not engage the Blues at all.

In order to understand index structure, consider a "one-step index" $H_j(n)$ defined by

$$H_j(n) = \frac{\sum_{b=1}^{B} \Pi_b^j (1-\rho_b)^n (1-\theta_b)^n (1-\phi_b)^n R_b \rho_b}{\sum_{b=1}^{B} \Pi_b^j (1-\rho_b)^n (1-\theta_b)^n (1-\phi_b)^n \{1 - \beta + \beta\theta_b\}}. \tag{14}$$

It is straightforward to establish the following:

(1) If $H_j(n)$ is decreasing in n, it then follows that $G_j(n) = H_j(n),\ n \in \mathbb{N}$. If this behaviour holds good for all Blues then Red's optimal shooting policy is quasi-myopic (a one-step look ahead rule). Here indices decrease through to Blue's destruction or departure and consequently the optimal index policy will tend to involve Red making frequent changes to the Blue targeted. In reality, set-up times would impose a penalty upon Red for such a policy. Glazebrook, Kirkbride and Ruiz (2005) propose modifications to indices which take account of switching penalties and/or times. Such modifications can be applied to all of the models discussed in this paper, although strict optimality is no longer achieved.

(2) If $H_j(n)$ is increasing in n, then the index $G_j(n)$ will take the form

$$\frac{\sum_{b=1}^{B} \Pi_b^j (1-\rho_b)^n (1-\theta_b)^n (1-\phi_b)^n R_b \rho_b \{1-\beta(1-\rho_b)(1-\theta_b)(1-\phi_b)\}^{-1}}{\sum_{b=1}^{B} \Pi_b^j (1-\rho_b)^n (1-\theta_b)^n (1-\phi_b)^n \{(1-\beta+\beta\theta_b)[1-\beta(1-\rho_b)(1-\theta_b)(1-\phi_b)]^{-1}\}}$$

and will itself be increasing in n. If this behaviour holds good for all Blues then Red, will in an optimal policy, persist in targeting individual Blues in turn until each is destroyed or withdraws. Note that in this event any disengagement by Red must either happen at $t = 0$ (no engagement at all) or immediately following the destruction or departure of one of the Blues.

Comments

(a) The one-step index $H_j(n)$ in (14) and the formula given in (2) above may both be thought of (somewhat crudely) as weighted averages (with respect to the posterior distribution) of a *return/exposure index*

$$R_b \rho_b \{1-\beta+\beta\theta_b\}^{-1}$$

for Blues of type b. This index is high when R_b and ρ_b are large and when θ_b is small. It is plainly such Blue types which Red should target early. Note the dependence of this quantity on θ_b. Plainly, Red should avoid targeting Blues with large associated θ-values as such engagements are high risk for her and her early demise will preempt the possibility of accumulating further returns. Note that in real circumstances the probability θ_b may be decreased by a reactive manoeuvre(s) by Red. Her success in such conflicts can depend upon sensor and communication properties which are only implicitly modelled here;

(b) While the above material has been presented for the case of a finite number of Blue types B in the interests of simplicity, it may be extended to cases in which type space is countable or continuous without difficulty. The underlying decision process continues to have a countable state space.

3.2 Model 2 – Red inflicts cumulative damage upon Blue

The distinctive feature of Model 2 is that the N Blues targeted by Red suffer cumulative damage during successive engagements. This is a step in the direction of shooting problems with targets whose characteristics evolve (degrade) dynamically. We shall here make the simplifying assumption that an engagement consists of a shot by Red at Blue j, say, followed by a retaliatory strike from the Blue targeted. Further, a more severely damaged Blue will be progressively less lethal to Red. Should a Blue's damage be sufficient to render it harmless, it is deemed to have been killed. To express this, we assume that each Blue can be in any one of D states, labelled $\{1, 2, \ldots, D\}$ and that this state is observable without error by Red. As state d runs from 1 to D it represents increasing degrees of damage with $D = \overline{\omega}_j$ corresponding to Blue's death. The Markovian law P^j determines how Blue j evolves to higher damage states under successive attacks from Red, while $\theta_j(d)$ is the probability that Blue j can kill Red with a shot when in damage state d, where $P^j_{dl} = 0,\ l < d$, and $\theta_j(D) = 0$. As in Model 1, a disengagement option of value R_d is always available to Red. We make the following natural assumptions:

Assumptions

(1) For all $j, \sum_{l=m}^{D} P^j_{dl}$ is increasing in d for each choice of $m \in \{d, d+1, \ldots, D\}$;

(2) For all j, $\theta_j(d)$ is decreasing in d with $\theta_j(D-1) > 0$.

Assumption (1) states that, following any engagement, Blue's new damage state is stochastically increasing in its old damage state. Assumption (2) states that Blue j becomes less lethal to Red as it is increasingly damaged. The condition $\theta_j(D-1) > 0$ allows us to make the choice $\beta = 1$ since condition (9) will then be met.

The general formulation of Section 2 (i)-(iv) can be adapted to this case as follows:

(i) State space Ω_j is $\{1, 2, \ldots, D\}$ with $D = \overline{\omega}_j$.

(iii) Should action a_j be chosen at t when $X_j(t) = d \in \{1, 2, \ldots, D-1\}$ then, following the resulting engagement between Red and Blue j a transition to $X_j(t+1)$ occurs according to Markovian law P_j where

$$P_j(d, l) = P(\text{engagement inconclusive, with Blue's damage } d \to l)$$

$$= P^j_{dl}\{1 - \theta_j(l)\},\ d \le l \le D-1;$$

$$P_j(d, D) = P(\text{Blue killed but not Red}) = P^j_{dD};$$

and

$$P_j(d, \omega_j) = P(\text{Red killed}) = \sum_{l=d}^{D-1} P^j_{dl}\theta_j(l).$$

The expected return from the engagement in (iii) above is given by

$$R_j(d) = \beta R_j P^j_{dD},\ d \in \{1, 2, \ldots, D-1\},$$

where we assume that the reward R_j is received when Blue j enters state D.

With the above specifications the index $G_j(d)$ appropriate for Blue j in damage state $d \in \{1, 2, \ldots, D-1\}$ may be easily computed. Red's optimal policy is again to target next the still alive Blue with maximal index up to the time of her death or the point at which all still alive Blues have index no greater than the disengagement return R_d. In the latter event, Red disengages.

To develop index structure, suppose that at time $t = 0$, Blue j is in damage state $d \in \{1, 2, \ldots, D-1\}$ and that Red engages with Blue j until one or the other is dead. Write T_d^j for the time at which the conflict ends and I_d^j for the indicator taking value 1 if the engagement ends with Blue j's death and 0 if it concludes with Red's demise. The key quantity

$$Z_d^j = E\left(\beta^{T_d^j} I_d^j\right)$$

may be interpreted as the probability that Red survives the engagement and satisfies the recursion

$$\begin{aligned} Z_d^j &= \beta P_{dD}^j + \beta \sum_{m=d}^{D-1} P_{dm}^j \{1 - \theta_j(m)\} Z_m^j \\ &= \beta \sum_{m=d}^{D} P_{dm}^j \{1 - \theta_j(m)\} Z_m^j, \end{aligned}$$

where we take $Z_D^j = 1$. A proof of the following result may be found in the on-line appendix. It makes use of a self-consistency result for Gittins indices first enunciated by Nash (1979).

Theorem 2

(i) The quantity Z_d^j is increasing in d, for each j, $1 \leq j \leq N$.

(ii) The index $G_j(d)$ for Blue j in damage state d is given by

$$G_j(d) = R_j Z_d^j (1 - Z_d^j)^{-1},\ d \in \{1, 2, \ldots, D-1\},\ 1 \leq j \leq N,$$

and is increasing in d.

Comments

(a) Note that the Blue index described in Theorem 2 has the feature that it is guaranteed to increase after each engagement through to the demise of one or other party. Further, it is the appropriate reward rate measure based on the assumption that Blues should be continuously engaged until killed. That our model should produce optimal policies of this character is natural since Blue's accumulating damage through his engagements not only brings his own death closer (Assumption (1)), but also makes him progressively less lethal to Red (Assumption (2)). Hence it is clear that Red should continue shooting at a partly damaged Blue and the index policy guarantees that this is so. Red will here only choose disengagement following the destruction of one of the Blues or at time $t = 0$.

(b) To see how the index $G_j(d)$ depends upon Blue j's lethality, consider two extreme cases. Suppose first that Blue j is lethal right up to its own destruction, namely

$$\theta_j(d) \cong 1,\ l \leq d \leq D-1.$$

It then follows that

$$Z_d^j \cong \beta P_{dD}^j$$

and hence that

$$G_j(d) \cong \beta R_j P_{dD}^j \{1-\beta P_{dD}^j\}^{-1}. \tag{15}$$

Note that the numerator in (15) is the expected return from a single shot (only) by Red at Blue j. In these circumstances, any fire from Red is a gamble that Blue j will be killed with a single shot. Suppose now that Blue j poses very little retaliatory threat to Red in that

$$\theta_j(d) \cong 0,\ 1 \leq d \leq D-1.$$

Consider the quantities $\{\tilde{Z}_d^j,\ 1 \leq d \leq D-1\}$ satisfying the recursions

$$\tilde{Z}_d^j = \beta P_{dD}^j + \beta \sum_{m=d}^{D} P_{dm}^j \tilde{Z}_m^j,\ 1 \leq d \leq D-1,\ \tilde{Z}_D^j = 1. \tag{16}$$

We now have

$$G_j(d) \cong R_j \tilde{Z}_d^j (1-\tilde{Z}_d^j)^{-1}, \tag{17}$$

a version of the index in Theorem 2 computed on the basis that Red will not be killed (other than by some external threat) in the conflict. Red's only concern here is the speed with which Blue j can be killed and the return R_j claimed. It follows straightforwardly from (16) that the index in (15) will be smaller than that in (17).

3.3 Model 3 – 'Shoot-look-shoot' for Red

The next goal is to give the reader some insight concerning the potential of our modelling/solution approach by introducing developments of Model 1 of considerable practical military import. The general scenario and Π_b^j, R_b, ρ_b and β are all as before. However, now suppose that after every shot by Red, the targeted Blue is inspected and categorised (with error) according to Blue target type and alive/dead. Write $\delta \in \{1, 2, \ldots, B\} \times \{\text{alive,dead}\}$ for a generic classification. We have that

$$P[\text{Blue judged to be } \delta | \text{Blue is alive of type } b] = \eta_{\delta b}$$
$$P[\text{Blue judged to be } \delta | \text{Blue is dead of type } b] = \eta_{\delta \bar{b}}$$

where $1 \leq b \leq B$. Also suppose that Red's vulnerability depends upon whether the targeted Blue is alive or dead. We use θ_b for the probability that Red is killed during an engagement in which she targets a Blue of type b who is still alive. This becomes $\overline{\theta}_b$ (typically less than

θ_b) if the targeted Blue is truly dead. We also use ϕ_b for the probability that an alive Blue of type b withdraws from the conflict following an unsuccessful shot by Red. This becomes $\overline{\phi}_b$ if the Blue concerned is truly dead (rendered ineffective). Blues rendered ineffective by Red will certainly wish to withdraw if they are able to do so. While Red may be uncertain regarding whether a Blue is alive or dead, she has no such uncertainty regarding a Blue's presence or absence.

Red now gathers information about the Blues she is facing through the series of engagements in a more complicated way than for Model 1. Index policies will remain optimal, but the index structure will be more complex and simple closed forms should certainly not be expected. Consider Blue target j with prior Π^j. At time t, if Red is still alive and Blue j still present then we require sufficient statistics from the history of Red's past engagements with targeted Blue j. These will determine Red's posterior distribution for this Blue. They are:

(a) the number of previous engagements targeting Blue j (n);

(b) the outcomes of Red's subsequent inspections ($\boldsymbol{\delta} = \{\delta_1, \delta_2, \ldots, \delta_n\}$).

We take these sufficient statistics as Blue j's state at t while Red is alive and Blue j present and write in vector notation $X_j(t) = (n, \boldsymbol{\delta})$. Note that we do not use the Blue identifier j in the data representation $(n, \boldsymbol{\delta})$ to ease the notational burden. Red's posterior probability, given the history $(n, \boldsymbol{\delta})$, that Blue j is of type b and is still alive is proportional to

$$\Pi_b^j (1-\rho_b)^n (1-\theta_b)^n (1-\phi_b)^n \left(\prod_{l=1}^{n} \eta_{\delta_l b} \right) \equiv \Pi_b^j P_b(n, \boldsymbol{\delta}) \equiv \Pi_b^j P_b\{X_j(t)\}. \tag{18}$$

Red's posterior probability, given this history, that Blue j is of type b but is now dead is proportional to

$$\Pi_b^j \sum_{k=1}^{n} (1-\rho_b)^{k-1} \rho_b (1-\theta_b)^k (1-\overline{\theta}_b)^{n-k} (1-\phi_b)^{k-1} (1-\overline{\phi}_b)^{n-k+1} \left(\prod_{l=1}^{k-1} \eta_{\delta_l b} \right) \left(\prod_{l=k}^{n} \eta_{\delta_l \overline{b}} \right)$$

$$\equiv \Pi_b^j \overline{P}_b(n, \boldsymbol{\delta}) \equiv \Pi_b^j \overline{P}_b\{X_j(t)\}, \tag{19}$$

as before. Hence, given the history summarised by $X_j(t)$, Red's posterior probabilities for Blue j are given by

$$P[\text{Blue } j \text{ is alive and of type } b | X_j(t)] = \frac{\Pi_b^j P_b\{X_j(t)\}}{\sum_{d=1}^{B} \Pi_d^j [P_d\{X_j(t)\} + \overline{P}_d\{X_j(t)\}]}, \quad 1 \le b \le B, \tag{20}$$

and

$$P[\text{Blue } j \text{ is dead and of type } b | X_j(t)] = \frac{\Pi_b^j \overline{P}_b\{X_j(t)\}}{\sum_{d=1}^{B} \Pi_d^j [P_d\{X_j(t)\} + \overline{P}_d\{X_j(t)\}]}, \quad 1 \le b \le B. \tag{21}$$

The formulation of Section 2 (i)-(iv) yields the following scheduling problem:

(i) State space Ω_j is the set of all possible histories $(n, \boldsymbol{\delta})$. Since in general under this model, Red can never be certain that Blue j has been killed, there is no state $\overline{\omega}_j$.

(iii) Suppose that action a_j is chosen at t when $X_j(t) = (n, \boldsymbol{\delta})$. If Red is not killed and Blue j does not withdraw, we have a state transition of the form

$$(n, \boldsymbol{\delta}) = X_j(t) \rightarrow X_j(t)(\delta) \equiv \{n+1, (\boldsymbol{\delta}, \delta)\}$$

with probability

$$\left(\sum_{b=1}^{B} \Pi_b^j [P_b\{X_j(t)\}\{(1-\rho_b)(1-\theta_b)(1-\phi_b)\eta_{\delta b} + \rho_b(1-\theta_b)(1-\overline{\phi}_b)\eta_{\delta\bar{b}}\} \right.$$
$$\left. + \overline{P}_b\{X_j(t)\}(1-\overline{\theta}_b)(1-\overline{\phi}_b)\eta_{\delta\bar{b}}] \right) \left(\sum_{b=1}^{B} \Pi_b^j [P_b\{X_j(t)\} + \overline{P}_b\{X_j(t)\}] \right)^{-1}. \quad (22)$$

If Red is killed then $X_j(t+1) = \omega_j$. This happens with probability

$$\frac{\sum_{b=1}^{B} \Pi_b^j [P_b\{X_j(t)\}\theta_b + \overline{P}_b\{X_j(t)\}\overline{\theta}_b]}{\sum_{b=1}^{B} \Pi_b^j [P_b\{X_j(t)\} + \overline{P}_b\{X_j(t)\}]}. \quad (23)$$

The expected return gained from a Blue kill in the engagement in (iii) above is given by

$$R_j(n, \boldsymbol{\delta}) = \frac{\sum_{b=1}^{B} \Pi_b^j P_b\{X_j(t)\} R_b \rho_b}{\sum_{b=1}^{B} \Pi_b^j \left[P_b\{X_j(t)\} + \overline{P}_b\{X_j(t)\}\right]}. \quad (24)$$

With the above specifications we may proceed to compute the index $G_j(n, \boldsymbol{\delta})$ appropriate for Blue j in state $(n, \boldsymbol{\delta})$. The authors recommend an adapted version of the "restart-in-$(n, \boldsymbol{\delta})$" approach to index computation proposed by Katehakis and Veinott (1987). See Glazebrook and Greatrix (1995) and refer to the first author for full details. Red will again target the Blue with maximal index until all indices are less than disengagement return R_d.

4 Modelling future opportunities – the nature of Red disengagement

Red will wish to disengage from any conflict if the value she places upon surviving "to fight another day" requires it. In Sections 2 and 3 we simply denoted this value R_d. However, the assignment of such a value implies that Red has some view of the future and (in particular) of the opportunities for securing enemy kills which it will bring. We now propose two possible models for the targets which Red will face as a process extended over time and discuss the implications for Red's decision-making, particularly with regard to disengagement.

4.1 The future as a sequence of intense Blue raids

Here Red's future will consist in confronting a sequence of discrete intense raids by Blue. These raids may be identical in character or, more generally, may be drawn at random from some finite *raid space* $\mathcal{R}$. Each member of $\mathcal{R}$ identifies the details of a single conflict type of the kind described in Section 2 and in Models 1–3 of Section 3. In particular, it will specify the character (and number) of Blues to be faced. Suppose that successive raids are drawn from $\mathcal{R}$ in an i.i.d. fashion according to the positive probabilities $\{\sigma_r, r \in \mathcal{R}\}$. Let $G(\mathcal{R})$ denote the maximal initial index (assumed finite) of *any* Blue appearing in *any* member of $\mathcal{R}$. Whenever Red is presented with a new Blue raid, she is able to determine its type $r \in \mathcal{R}$ and must decide how to shoot during the raid and when to disengage from shooting. Suppose that the times between successive raids (i.e. between Red's disengagement from one raid and the commencement of the next) form a sequence of i.i.d. positive-valued random variables whose distribution is that of the random variable T. Write

$$m = E\left(\beta^T\right).$$

It will simplify matters if we suppose that β and m are both (strictly) less than one for the purposes of this discussion. Consider a scenario in which the first raid faced by Red is at time zero. Denote Red's maximised expected total return from time zero but before the determination of the type of the first raid by $V(m)$. The stationarity of the model implies that in each raid, the (undiscounted) value which Red should place on disengagement is $mV(m) = R_d$. Plainly, from the discussion and results of Sections 2 and 3, in any raid Red will optimally shoot at Blues according to an appropriate index policy and will only disengage when all remaining Blue targets have indices which are less than $mV(m)$.

In order to develop ideas we shall need the following notation: consider a policy for Red during a raid of type $r \in \mathcal{R}$ in which she shoots according to an index policy and disengages when all Blue indices are less than $x \in \mathbb{R}^+$. The best (reward maximising) level of x is the *optimal disengagement level.* If the raid begins at 0, write $\tau_r(x)$ for the time of Red's death or disengagement, whichever comes first. Also use $I_r(x)$ for the indicator which is 0 if the raid ends with Red's death and is 1 otherwise. Finally $R_r(x)$ is the expected (discounted) return gained from Blue kills during the raid. From the above discussion we may assert that

$$\begin{aligned} V(m) &= \sum_{r\in\mathcal{R}} \sigma_r \left[R_r\{mV(m)\} + E\left[\beta^{\tau_r\{mV(m)\}} I_r\{mV(m)\}\right] mV(m)\right] \\ &\geq \sum_{r\in\mathcal{R}} \sigma_r \left[R_r(x) + E\left\{\beta^{\tau_r(x)} I_r(x)\right\} mV(m)\right],\ x \in \mathbb{R}^+. \end{aligned} \tag{25}$$

The following result is a consequence of (25), the foregoing discussion and the theory of Gittins indices. A proof may be found in the on-line appendix.

Theorem 3

(i) Red's maximised total expected return is given by

$$V(m) = \max_{x\in\mathbb{R}^+} \left\{\sum_{r\in\mathcal{R}} \sigma_r R_r(x)\right\} \left\{\sum_{r\in\mathcal{R}} \sigma_r \left[1 - mE\left\{\beta^{\tau_r(x)} I_r(x)\right\}\right]\right\}^{-1}, \tag{26}$$

and is increasing in m with the maximum in (26) *achieved at the optimum disengagement level $x = mV(m)$;*

(ii) Let $\{T_n, n \in \mathbb{N}\}$ be a sequence of non-negative-valued r.v.'s with $m_n = E\left(\beta^{T_n}\right) \uparrow 1,\ n \to \infty$. It will follow that $\{V(m_n), n \in \mathbb{N}\}$ and $\{m_n V(m_n), n \in \mathbb{N}\}$ are both increasing sequences with

$$\lim_{n\to\infty} V(m_n) = \lim_{n\to\infty} m_n V(m_n) = G(\mathcal{R}). \tag{27}$$

Comments

In Theorem 3 and the following observations we fix all aspects of the model, save only the choice of T and the consequential value of m. If $m \cong 0$ then it must be that the times between successive raids are large and the gains from future raids consequently heavily discounted. When this is the case

$$V(m) \cong \max_{x\in\mathbb{R}^+} \sum_{r\in\mathcal{R}} \sigma_r R_r(x), \tag{28}$$

where the maximisation in (28) is of the expected return from a single raid. It is obvious (and, indeed, is a consequence of Theorem 3(i)) that the maximum in (28) is achieved at $x = 0$. Hence Red will be reluctant to disengage from any such conflict in the absence of future value.

In contrast, if $m \approx 1$ then from Theorem 3, the maximum in (26) is achieved at $mV(m) \approx G(\mathcal{R})$. Hence when returns from future raids are subject to light discounting, Red should be very selective about the Blues she targets. In the limit as m approaches 1, Red disengages as soon as there are no available targets of index value at least equal to the maximal initial value $G(\mathcal{R})$. Theorem 3 makes formal the notion that as the frequency of raids (measured by m) increases, Red becomes progressively more selective about the Blues she targets and disengages earlier from every raid type.

Note finally that for any fixed m, $V(m)$ may be computed by a form of DP value iteration. The following result may be established using Theorem 3 together with arguments based on monotone mappings. A proof may be found in the on-line appendix. We use f^n for an n-fold application of function f.

Lemma 4 *The function $f : \mathbb{R}^+ \to \mathbb{R}^+$, defined by*

$$f(x) = \sum_{r\in\mathcal{R}} \sigma_r \left[R_r(mx) + E\left\{\beta^{\tau_r(mx)} I_r(mx)\right\} mx\right],\ x \in \mathbb{R}^+$$

is such that

$$\lim_{n\to\infty} f^n(0) = V(m).$$

4.2 Poisson arrivals of multiple Blue types

In contrast to the sporadic periods of intense activity envisaged in (4.1), we now suppose that Red faces a Poisson stream of Blue targets over time. As we shall see, the insights we derive regarding disengagement are qualitatively similar to those in the preceding subsection. Suppose now that each Blue target belongs to one of C classes with distinct members of the same class having identical characteristics, but experiencing independent outcomes. Individual Blue targets are as in Section 2 and in Models 1–3 of Section 3. Blues from class $c \in \mathcal{C} \equiv \{1, 2, \ldots, C\}$ arrive according to a Poisson stream with rate λ_c, with streams from distinct classes independent. We write $\Lambda = \sum_{c \in \mathcal{C}} \lambda_c$ for the total arrival rate and $G(\mathcal{C})$ for the maximal initial index from the C classes.

When there are Blues present, Red may choose to shoot at one of them (which will take a single unit of time) or she may choose to disengage and wait for further targets to arrive. Should Red choose to disengage then she must wait an $\exp(\Lambda)$ period of time before further Blues arrive. At that point she may *either* resume shooting *or* remain disengaged. Note that in the former case and under an optimal policy Red will never again shoot at any Blues which were present at an earlier decision to disengage. Write $V(\boldsymbol{\lambda}, \beta)$ for the expected return achieved by Red up to her death when adopting an optimal policy for shooting/disengagement and when no targets are present at time zero. We use

$$W(\boldsymbol{\lambda}, \beta) \equiv (\Lambda - \ln \beta)(\Lambda)^{-1} V(\boldsymbol{\lambda}, \beta)$$

for the equivalent quantity when zero is taken as the time of arrival of the first Blue target. Exploiting and extending the observation in Section 2 that our core models (without arrivals) may be regarded as (semi-Markov) multi-armed bandits, Red's problem may be modelled as a semi-Markov branching bandit problem in which Red's disengagement option has a fixed index value equal to $V(\boldsymbol{\lambda}, \beta)$. While it is true that the indices solving Red's shooting/disengagement problem will now in general depend upon the vector $\boldsymbol{\lambda}$ of arrival rates, the target ordering implied by the indices is rarely different from that for the equivalent closed case $\boldsymbol{\lambda} = \mathbf{0}$. See Fay and Glazebrook (1992) who discuss precisely the closeness to optimality of a so-called "no arrivals" index heuristic. Hence choosing between the Blue targets currently present on the basis of indices of the kind described in Sections 2 and 3 will be close to optimal for Red. Note also that the value of $G(\mathcal{C})$ is *not* $\boldsymbol{\lambda}$-dependent.

Consider a situation in which a single Blue from class c is present at time zero. If Red shoots optimally from time zero but disengages as soon as all indices are less than $x \in \mathbb{R}^+$, then $R_c(x)$ is the expected return from Blue kills prior to first disengagement, $\tau_c(x)$ is the time of Red's death or first disengagement (whichever comes first) and $I_c(x)$ is the indicator which is 0 if Red is dead at $\tau_c(x)$ and which is 1 otherwise. The argument yielding the following result is similar to that which gave Theorem 3. In Theorem 5 we write $\bar{c}$ for (one of) the class(es) achieving $G(\mathcal{C})$.

Theorem 5

(i) Red's maximised total expected return is given by

$$V(\boldsymbol{\lambda}, \beta) = \Lambda(\Lambda - \ln \beta)^{-1} \max_{x \in \mathbb{R}^+} \left\{ \sum_{c=0}^{C} \lambda_c R_c(x) \right\} \left(\sum_{c=0}^{C} \lambda_c \left[1 - \Lambda(\Lambda - \ln \beta)^{-1} E\left\{ \beta^{\tau_c(x)} I_c(x) \right\} \right] \right)^{-1},$$

with the maximum achieved at the optimum disengagement level $x = V(\boldsymbol{\lambda}, \beta)$;

(ii) $V(\boldsymbol{\lambda}, \beta)$ is increasing componentwise in $\boldsymbol{\lambda}$. If $\lambda_{\bar{c}} \to \infty$ (with other components of $\boldsymbol{\lambda}$ fixed) then $V(\boldsymbol{\lambda}, \beta) \to G(\mathcal{C})$.

Comments

(a) Similar comments to those following Theorem 3 apply. If Λ is close to 0 (Blue targets arrive sporadically) then Red should engage almost all of them. If, however, there are copious supplies of targets from class $\bar{c}$ then Red should be very selective and only engage those whose indices are no less than $G(\mathcal{C})$.

(b) It is certainly possible to model futures for Red other than those in (4.1) and (4.2). These include a hybrid of the above models of compound Poisson type in which intense Blue raids arrive according to a Poisson process.

5 Numerical Study

We report on the outcome of a numerical study whose aim is to give the reader some sense of the reward advantages to be gained by the adoption by Red of an optimal (index) policy for shooting and also to quantify the value of disengagement in a range of scenarios. Below are reported results for three problem sets (1-3) chosen to represent a range of operational alternatives.

For each set, we report first on the rewards gained by Red under a range of policies for a one-off conflict with Blue which has no disengagement option (equivalently, $R_d = 0$). All cases studied are instances of Model 1 in (3.1) with $N = 10$ (ten Blue targets) and $B = 5$ (five Blue types). The discount rate β (Red survival probability per unit of time) is taken to be 0.95 throughout. Table 1 contains details of the Blue types for each problem. Note that $\phi_b = 0, \ 1 \leq b \leq 5$, namely that in these examples Blues do not withdraw under fire.

	PROBLEM SET 1			PROBLEM SET 2			PROBLEM SET 3		
b	ρ_b	θ_b	R_b	ρ_b	θ_b	R_b	ρ_b	θ_b	R_b
1	0.8	0.10	60	0.9	0.10	100	0.9	0.2	50
2	0.7	0.15	70	0.8	0.05	125	0.7	0.3	125
3	0.6	0.08	80	0.5	0.01	250	0.5	0.4	150
4	0.5	0.05	90	0.6	0.20	750	0.3	0.5	500
5	0.4	0.40	200	0.4	0.40	1000	0.1	0.6	1000

Table 1: Details of the Blue types for each problem set

The simulation study of one-off conflicts consists of 18×10^6 runs – with 10^6 runs being conducted for each of six different policies for Red for each of the three problem sets. For each of the runs for each problem set a prior for Red is set as follows: the ten Blue targets are grouped into five pairs. One of the i^{th} pair has an assigned prior probability of 0.75 for Blue type i while the other has an assigned probability of 0.50, $1 \leq i \leq 5$. The remaining prior probabilities are obtained by drawing independently from a $U(0, 1)$ distribution and normalising appropriately. For each individual run, actual Blue types are determined by drawing from the appropriate prior. The six shooting policies for Red are as follows:

(I) **Index** (IN) – This is the policy which maximises the expected return earned by Red before her death;

(II) **Myopic** (MY) – Here Red's policy is to shoot next at whichever Blue is still alive and offers her the highest expected one-stage return. Hence, the quantity in (13) is used as a calibrating index for Blue j;

(III) **Survival** (SU) – Here Red's next shot is targeted in such a way as to give the largest probability of surviving the engagement. Hence, the quantity $1 - P_j(n, \omega_j)$ (see (12)) is used as a calibrating index for Blue j;

(IV) **Exhaustive** (EX) – Here Red adopts the best policy among those in which she shoots continuously at each Blue targeted until either party to the engagement is killed. This optimisation problem calls for a simple ordering of the Blue targets and may also be formulated as a multi-armed bandit. Red should shoot at Blues in the order of decreasing values (i.e., largest first) of the quantities

$$\frac{\sum_{b=1}^{B} \Pi_b^j R_b \rho_b \{1-\beta(1-\rho_b)(1-\theta_b)\}^{-1}}{\sum_{b=1}^{B} \Pi_b^j (1-\beta+\beta\theta_b)\{1-\beta(1-\rho_b)(1-\theta_b)\}^{-1}}, \quad 1 \leq j \leq 10;$$

(V) **Random** (RA) – At each stage, Red chooses between the still-alive Blues at random, with all Blue targets equally likely;

(VI) **Round Robin** (RR) – Red cycles around the Blue targets (which are still alive) in numerical order. The first target is chosen at random.

	Policy	Mean	LQ	Med	UQ	NBK
PROBLEM SET 1	IN	299.42	149.82	295.57	427.78	4.72
	MY	206.71	0.00	190.00	321.77	2.21
	SU	298.91	157.70	286.65	427.05	4.76
	EX	299.10	149.82	295.17	427.27	4.71
	RA	249.60	85.50	215.39	372.80	3.50
	RR	249.94	85.50	215.93	373.71	3.51
PROBLEM SET 2	IN	1126.04	621.60	1026.55	1587.03	4.29
	MY	931.25	0.00	950.00	1433.43	2.17
	SU	1061.75	488.54	1018.21	1473.53	4.97
	EX	1125.96	621.60	1026.21	1586.25	4.29
	RA	973.66	317.30	919.59	1446.61	3.54
	RR	978.83	350.31	912.90	1447.20	3.58
PROBLEM SET 3	IN	259.10	0.00	118.75	475.00	0.91
	MY	258.19	0.00	47.50	475.00	0.80
	SU	218.45	47.50	160.31	273.13	1.99
	EX	258.13	0.00	118.75	475.00	0.91
	RA	224.25	0.00	118.75	327.16	1.18
	RR	225.17	0.00	118.75	333.09	1.23

Table 2: Summary of Red's returns and numbers of Blues killed using six different shooting policies for Red for three problem sets

Table 2 contains summaries of the 18×10^6 runs conducted. For each choice of policy/problem set it gives a statistical summary of the 10^6 returns earned by Red and records the mean return, the lower quartile (LQ), median (Med) and upper quartile (UQ). The final column records the mean number of Blues killed (NBK). As predicted by the theory, IN dominates the other policies with respect to the mean return gained. Note also that the problem set-ups are such that in every case IN operates very similarly to at least one other policy. For problem set 1 the policies (IN, SU, EX) are very close; in set 2 this is true of (IN, EX) and in set 3 of (IN, MY, EX). The closeness of IN and EX is not a universal feature of Model 1 and is present here because our priors reflect reasonably strong prior beliefs on Red's part as to which type each Blue target is. The very poor performance of MY for sets 1 and 2 is rooted in its indifference to the issue of Red's vulnerability. In set 3, SU is overly cautious and leads Red to overlook high gains in favour of survival. In these cases Red would be better off choosing Blue targets at random (or in a round robin fashion). Unsurprisingly, policy SU dominates the final column (NBK) and operates in such a way as to favour numbers of Blue kills rather than the values thereof.

The second part of the study is an exploration of the value to Red of disengagement. To progress we continue to use the problem sets above but now suppose that Red faces a Blue target process consisting of a sequence of identical conflicts in the manner of subsection (4.1). For a range of values of m between 0.25 and 0.95, an estimate of Red's optimal total expected return from targeting/disengagement, namely $\hat{V}(m)$, is obtained using a hybrid approach involving the value iteration of Lemma 4 and simulation. These values are then deployed in a simulation study which estimates and compares Red's returns when shooting and disengaging from each conflict optimally with those obtained when Red shoots optimally but never disengages and only proceeds to later conflicts if she survives earlier ones.

	m	$\hat{V}(m)$	Mean	LQ	Med	UQ	MND
PROBLEM SET 1	0.25	301.52	301.45	149.82	295.46	427.80	301.23
	0.50	303.64	303.99	149.82	295.90	428.58	303.52
	0.75	308.15	308.19	149.82	295.57	435.32	305.88
	0.80	310.03	310.04	149.82	295.57	437.16	305.90
	0.85	314.06	313.88	149.82	295.58	450.24	306.68
	0.90	320.63	320.57	149.82	296.83	463.67	307.16
	0.95	340.83	340.65	157.70	327.40	499.15	307.51
PROBLEM SET 2	0.25	1134.67	1134.06	621.60	1025.57	1590.71	1133.88
	0.50	1143.20	1143.42	621.60	1026.94	1597.20	1142.58
	0.75	1151.98	1152.84	621.60	1028.10	1601.79	1151.83
	0.80	1154.01	1154.57	621.60	1028.36	1603.71	1153.34
	0.85	1167.15	1168.13	621.60	1051.95	1654.72	1155.26
	0.90	1193.74	1194.11	621.60	1087.46	1696.47	1157.58
	0.95	1234.78	1234.70	624.39	1138.92	1774.00	1158.69

Table 3: Summary of Red's returns using both optimal disengagement (Mean,LQ,Med,UQ) and no disengagement (MND) for two problem sets

Table 3 contains summaries of the outcomes of 14×10^6 runs conducted for problem sets 1 and 2. For each choice of m-value/problem set it gives the value of $\hat{V}(m)$ together with a statistical summary of the 10^6 returns earned by Red when deploying optimal disengagement (i.e. disengage when all current indices fall below $m\hat{V}(m)$). This summary includes the mean return (which also estimates $V(m)$), the lower quartile (LQ), median (Med) and upper quartile (UQ). The final column gives an estimate of the best mean return achievable by Red if she never disengages (MND). When comparing values of $\hat{V}(m)$ (or Mean) with those of MND it is clear that for these problem sets the benefits of disengagement increase with m and can become considerable if $m \cong 1$. This is usually, but not always the case. In an equivalent study for problem set 3, $\hat{V}(0.95)$ was just 0.1% larger than $\hat{V}(0.25)$. Here, Red's death comes quickly, conflicts are short and there is little opportunity for disengagement. Even when $m = 0.95$, if Red disengages optimally she only exercises that option in 1.8% of conflicts. For problem set 3, then, it is the choice of Blue targets which is the critical issue (see Table 2).

Acknowledgements

The authors are grateful to an Associate Editor and two referees whose thoughtful comments prompted futher work and considerable strengthening of the paper. The first three authors acknowledge the support provided by the Engineering and Physical Sciences Council, including that through grant GR/S45188/01. The last two authors acknowledge the support provided by the Naval Postgraduate School Center for Defense Technology and Education in the Military Services (CDTEMS).

References

Barkdoll, T. C., Gaver, D. P., Glazebrook, K. D., Jacobs, P. A. & Posadas, S. (2002), 'Suppression of Enemy Air Defences (SEAD) as an Information Duel', *Nav. Res. Logist.* **49**, 723–742.

Bertsimas, D. & Niño-Mora, J. (1996), 'Conservation laws, extended polymatroids and multi-armed bandit problems: a polyhedral approach to indexable systems', *Math. Oper. Res.* **21**, 257–306.

Crosbie, J. H. & Glazebrook, K. D. (2000*a*), 'Evaluating policies for generalized bandits via a notion of duality', *J. Appl. Probab.* **37**, 540–546.

Crosbie, J. H. & Glazebrook, K. D. (2000*b*), 'Index policies and a novel performance space structure for a class of generalised branching bandit problems', *Math. Oper. Res.* **25**, 281–297.

Dumitriu, I., Tetali, P. & Winkler, P. (2003), 'On playing golf with two balls', *SIAM J. Discrete Math.* **16**, 604–615.

Fay, N. A. & Glazebrook, K. D. (1992), 'On a "no arrivals" heuristic for single-machine stochastic scheduling', *Oper. Res.* **40**, 168–177.

Fay, N. A. & Walrand, J. C. (1991), 'On approximately index strategies for generalized arm problems', *J. Appl. Probab.* **28**, 602–612.

Gaver, D. P., Glazebrook, K. D. & Pilnick, S. E. (1991), 'Optimal sequential replenishment of ships during combat', *Nav. Res. Logist.* **38**, 637–668.

Gittins, J. C. (1979), 'Bandit processes and dynamic allocation indices (with discussion)', *J. Roy. Statist. Soc.* **B41**, 148–177.

Gittins, J. C. (1989), *Multi-armed Bandit Allocation Indices*, Wiley, Chichester.

Gittins, J. C. & Jones, D. M. (1974), A dynamic allocation index for the sequential design of experiments, *in* 'Progress in Statistics', J. Gani & I. Vince, eds, North-Holland, Amsterdam, pp. 241–266.

Glazebrook, K. D., Ansell, P. S., Dunn, R. T. & Lumley, R. R. (2004), 'On the optimal allocation of service to impatient tasks', *J. Appl. Probab.* **41**, 51–72.

Glazebrook, K. D. & Greatrix, S. (1995), 'On transforming an index for generalised bandit problems', *J. Appl. Probab.* **32**, 168–182.

Glazebrook, K. D., Kirkbride, C. & Ruiz, D. (2005), 'Some families of indexable restless bandit problems', *submitted for publication* .

Glazebrook, K. D. & Washburn, A. (2004), 'Shoot-look-shoot: A review and extension', *Oper. Res.* **52**, 454–463.

Katehakis, M. N. & Veinott, A. F. (1987), 'The multi-armed bandit problem - decomposition and computation', *Math. Oper. Res.* **12**, 262–268.

Katta, A. & Sethuraman, J. (2004), 'A note on bandits with a twist', *SIAM J. Discrete Math.* **18**, 110–113.

Manor, G. & Kress, M. (1997), 'Optimality of the greedy shooting strategy in the presence of incomplete damage information', *Nav. Res. Logist.* **44**, 613–622.

Mitchell, H. M. (2003), PhD thesis, Newcastle University, Newcastle upon Tyne, UK.

Nash, P. (1979), PhD thesis, Cambridge University, UK.

Nash, P. (1980), 'A generalised bandit problem', *J. Roy. Statist. Soc.* **B42**, 165–169.

Papadimitriou, C. H. & Tsitsiklis, J. N. (1999), 'The complexity of optimal queueing network control', *Math. Oper. Res.* **24**, 293–305.

Puterman, M. L. (1994), *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York.

Robinson, D. (1982), 'Algorithms for evaluating the dynamic allocation index', *Oper. Res. Lett.* **1**, 72–74.

Ross, S. M. (1970), *Applied probability models with optimization applications*, Holden-Day, San Francisco.

U.S. Marine Corps (2001), 'Suppression of Enemy Air Defenses (SEAD)', *MCWP.* 3–22·2.

Weber, R. R. (1992), 'On the Gittins index for multi-armed bandits', *Ann. Appl. Probab.* **2**, 1024–1035.

Whittle, P. (1980), 'Multi-armed bandits and the Gittins index', *J. Roy. Statist. Soc.* **B42**, 143–149.

Whittle, P. (1988), 'Restless bandits: activity allocation in a changing world', *J. Appl. Probab.* **A25**, 287–398.

INITIAL DISTRIBUTION LIST

1. Research Office (Code 09)..1
   Naval Postgraduate School
   Monterey, CA 93943-5000

2. Dudley Knox Library (Code 013)..2
   Naval Postgraduate School
   Monterey, CA 93943-5002

3. Defense Technical Information Center..2
   8725 John J. Kingman Rd., STE 0944
   Ft. Belvoir, VA 22060-6218

4. Richard Mastowski (Editorial Assistant)..2
   Graduate School of Operational and Information Sciences (GSOIS)
   Naval Postgraduate School
   Monterey, CA 93943-5219

5. Professor Kevin D. Glazebrook...1
   Department of Management Science
   Management School
   University of Lancaster
   Lancaster, LA1 4YX
   UK

6. Dr. C. Kirkbride...1
   Department of Management Science
   Management School
   University of Lancaster
   Lancaster, LA1 4YX
   UK

7. Dr. H.M. Mitchell 1
   Department of Mathematics and Statistics
   University of Newcastle
   Newcastle-upon-Tyne
   NE1 7RU
   UK

8. Professor Peter Denning .... Electronic copy
   pjd@nps.edu

9. Professor Moshe Kress .... Electronic copy
   mkress@nps.edu

10. Professor Kyle Lin .... Electronic copy
kylin@nps.edu

11. Professor Roberto Szechtman .... Electronic copy
rszecht@nps.edu

12. Distinguished Professor Donald P. Gaver 3
Department of Operations Research
Naval Postgraduate School
Monterey, CA 93943-5219

13. Professor Patricia A. Jacobs..2
Department of Operations Research
Naval Postgraduate School
Monterey, CA 93943-5219

14. Professor Frank Shoup .... Electronic copy
feshoup@nps.edu

15. Distinguished Professor Emeritus... Electronic copy
awashburn@nps.edu

16. Professor Michael Armstrong .... Electronic copy
mjarmstrong@nps.edu

17. MDA Working Group .... Electronic copy
MDAWG@nps.edu